

Isoform reconstruction using short RNA-Seq reads by maximum likelihood is NP-hard

Tianyang Li

Rui Jiang

Xuegong Zhang*

MOE Key Laboratory of Bioinformatics
and Bioinformatics Division,
TNLIST/Department of Automation,
Tsinghua University,
Beijing, China, 100084

*Correspondence: zhangxg@tsinghua.edu.cn

Abstract—Maximum likelihood is a popular technique for isoform reconstruction. Here, we show that isoform reconstruction using short RNA-Seq reads by maximum likelihood is NP-hard.

I. INTRODUCTION

Isoform reconstruction is a key step in RNA-Seq analysis. Tools such as CEM [1], iReckon [2], NSMAP [3], and Montebello [4] use maximum likelihood for isoform reconstruction. The maximum likelihood approach has been observed to be computationally expensive. Here, we show that isoform reconstruction using short RNA-Seq reads by maximum likelihood is NP-hard.

II. RESULTS

A Poisson mixture model [5]–[7] is used for isoform reconstruction. We represent a gene as a directed acyclic graph $G = (V, E)$ where each vertex in G represents an exon, and a path in G represents an isoform of this gene [8], [9]. In the model [7], the likelihood of observing N_s at each read 3' end location equivalence class s is

$$\prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{N_s}}{N_s!} \quad (1)$$

Here S is the set of all read 3' end location equivalence classes, and

$$\lambda_s = \sum_{i \in I_s} a_{is} \theta_i \quad (2)$$

where I_s is the set of isoforms compatible with s , θ_i is isoform i 's expression level, and $a_{is} > 0$ is the sampling rate for read 3' end location equivalence class s on isoform i [6].

To reconstruct isoforms, we seek to maximize the likelihood (1) with a set I consisting of isoforms, and each isoform's expression level θ_i , $i \in I$. In order to explain all observed reads, we must be able to align each read to at least one isoform in I . However, because there are a large number of possible isoforms, and it is generally believed that a gene only has a small number of highly expressed isoforms, we

instead try to find I and θ_i , $i \in I$ that maximize the following penalized likelihood

$$e^{-K\|I\|_0} \prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{N_s}}{N_s!} \quad (3)$$

where $\|I\|_0$ is the number of $\theta_i > 0$, $i \in I$, and $K > 0$ is a real constant. Note that setting $k = \frac{1}{2} \log(\sum_{s \in S} N_s)$ is equivalent to using the Bayesian information criterion [10] for an equivalent multinomial model [7], [11].

To show the hardness of isoform reconstruction by maximizing the penalized likelihood (3), we consider the following decision problem

M-ISOFORM.

INSTANCE: A set of reads aligned to a gene where the read count at a read 3' end location equivalence class s the read count is N_s .

QUESTION: Does there exist an isoform set I with at most m isoforms such that $\prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{N_s}}{N_s!} \geq p$?

Theorem 1. M-ISOFORM is NP-complete.

III. DISCUSSION

We can avoid computationally determining a gene's isoform set if laboratory protocols can be used to find the gene's existing isoforms. It is possible to use methods such as paired-end tag sequencing [12], and single-molecule sequencing [13] to determine a gene's isoforms. We expect these and related technologies to mature in the foreseeable future, reducing the demand for computational resources.

Haplotype inference [14], [15], a similar problem, is also known to be NP-hard [16]. We believe that haplotype inference will also benefit from technologies offering longer sequencing reads.

IV. PROOF

The proof borrows ideas from [17], where a network flow approach is used for isoform reconstruction. To show that M-ISOFORM is NP-complete, we reduce 3-PARTITION [18], [19], a strongly NP-complete problem, to M-ISOFORM. We

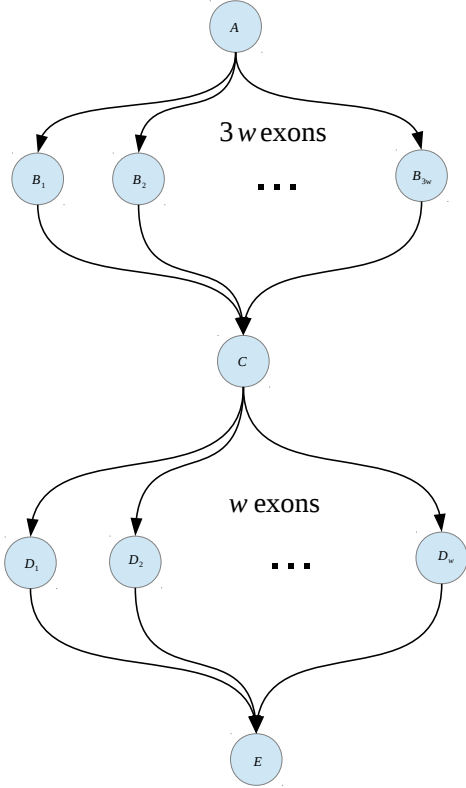


Fig. 1: A gene structure for an instance of 3-PARTITION

use an approach similar to the one used in [20], where flows are split into paths.

Proof of Theorem 1: 3-PARTITION is stated in [18] as follows

3-PARTITION.

INSTANCE: Set X of $3w$ elements, a bound $Y \in \mathbb{Z}^+$, and a size $u(x) \in \mathbb{Z}^+$ for each $x \in X$ such that $\frac{Y}{4} < u(x) < \frac{Y}{2}$ and such that $\sum_{x \in X} u(x) = wY$.

QUESTION: Can X be partitioned into w disjoint sets X_1, X_2, \dots, X_w such that, for $1 \leq i \leq w$, $\sum_{x \in X_i} u(x) = Y$ (note that each X_i must therefore contain exactly three elements from X)?

For an instance of 3-PARTITION, we create a gene with structure as shown in Figure 1. Let $E \in \mathbb{Z}^+$ be a fixed constant, we make each exon R bp long, and each read $R+1$ bp long. At each bp in exon A , for $1 \leq i \leq 3w$ we have $u(x_i)$ reads starting at this location going to exon B_i , thus there are $\sum_{1 \leq i \leq 3w} u(x_i)$ reads starting at this location. For $1 \leq i \leq 3w$, at each bp in exon B_i , we have $u(x_i)$ reads starting at this location going to exon C . At each bp in exon C , for $1 \leq i \leq w$ we have Y reads starting at this location going to exon D_i , thus there are wY reads starting at this location. For $1 \leq i \leq w$, at each bp in exon D_i , we have Y reads starting at this location going to exon E . We will show

that this instance of 3-PARTITION has a solution if and only if there exists an isoform set I with at most $3z$ isoforms such that $\prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{N_s}}{N_s!} \geq \prod_{s \in S} \frac{e^{-N_s} N_s^{N_s}}{N_s!}$.

It is easy to see that $\prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{N_s}}{N_s!} \geq \prod_{s \in S} \frac{e^{-N_s} N_s^{N_s}}{N_s!}$ if and only if $\forall s \in S, \lambda_s = N_s$.

If we have a solution to this instance of 3-PARTITION, it is easy to verify that an isoform set with $3z$ isoforms where $\theta_i = u(x_i)$, $1 \leq i \leq 3z$, and isoform i consists of exons A, B_i, C, D_j (j satisfies $x_i \in A_j$), and E is a solution to this particular instance of M-ISOFORM.

If we have a solution to this particular instance of M-ISOFORM, we show that there also exists a solution to the instance of 3-PARTITION. In this case, the isoform set must have exactly $3z$ isoforms, because at least $3z$ isoforms are required to explain all the reads. Thus, for $1 \leq i \leq 3z$ we have $\theta_i = u(x_i)$. Because we also have $\forall s \in S, \lambda_s = N_s$, and $\forall x \in X, \frac{Y}{4} < u(x) < \frac{Y}{2}$ we can see that for $1 \leq j \leq w$ exon D_j has exactly three isoforms passing through it, and \sum_i passes through D_j $\theta_i = Y$. Therefore, we have a solution to the instance of 3-PARTITION.

It is easy to see that M-ISOFORM is in NP if we use the real RAM model [21]. Because 3-PARTITION is strongly NP-complete [18], [19], we conclude that M-ISOFORM is NP-complete. ■

ACKNOWLEDGMENTS

We thank Feng Zeng for insightful discussions.

REFERENCES

- [1] W. Li and T. Jiang, "Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads," *Bioinformatics*, vol. 28, no. 22, pp. 2914–2921, 2012. [Online]. Available: <http://bioinformatics.oxfordjournals.org/content/28/22/2914.abstract>
- [2] A. M. Mezlini, E. J. Smith, M. Fiume, O. Buske, G. Savich, S. Shah, S. Aparicion, D. Chiang, A. Goldenberg, and M. Brudno, "iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data," *Genome Research*, 2012. [Online]. Available: <http://genome.cshlp.org/content/early/2012/11/29/gr.142232.112.abstract>
- [3] Z. Xia, J. Wen, C.-C. Chang, and X. Zhou, "NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq," *BMC Bioinformatics*, vol. 12, no. 1, p. 162, 2011. [Online]. Available: <http://www.biomedcentral.com/1471-2105/12/162>
- [4] D. Hiller and W. Wong, "Simultaneous Isoform Discovery and Quantification from RNA-Seq," *Statistics in Biosciences*, pp. 1–19, 2012. [Online]. Available: <http://dx.doi.org/10.1007/s12561-012-9069-2>
- [5] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in rna-seq," *Bioinformatics*, vol. 25, no. 8, pp. 1026–1032, 2009. [Online]. Available: <http://bioinformatics.oxfordjournals.org/content/25/8/1026.abstract>
- [6] J. Salzman, H. Jiang, and W. H. Wong, "Statistical Modeling of RNA-Seq Data," *ArXiv e-prints*, Jun. 2011.
- [7] L. Pachter, "Models for transcript quantification from RNA-Seq," *ArXiv e-prints*, Apr. 2011.
- [8] S. Heber, M. Alekseyev, S.-H. Sze, H. Tang, and P. A. Pevzner, "Splicing graphs and EST assembly problem," *Bioinformatics*, vol. 18, no. suppl 1, pp. S181–S188, 2002. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S181.abstract
- [9] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, and et al., "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, no. 5, pp. 503–510, 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20436462>

- [10] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. [Online]. Available: <http://projecteuclid.org/euclid.aos/1176344136>
- [11] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20436464>
- [12] M. J. Fullwood, C.-L. Wei, E. T. Liu, and Y. Ruan, "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses," *Genome Research*, vol. 19, no. 4, pp. 521–532, 2009. [Online]. Available: <http://genome.cshlp.org/content/19/4/521.abstract>
- [13] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and et al., "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nature Biotechnology*, vol. 30, no. 7, pp. 693–700, 2012. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/22750884>
- [14] L. M. Li, J. H. Kim, and M. S. Waterman, "Haplotype reconstruction from SNP alignment." *Journal of Computational Biology*, vol. 11, no. 2-3, pp. 505–516, 2004. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15285905>
- [15] E. Xing, R. Sharan, and M. I. Jordan, "Bayesian haplotype inference via the Dirichlet process," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 111–. [Online]. Available: <http://doi.acm.org/10.1145/1015330.1015423>
- [16] R. Sharan, B. Halldorsson, and S. Istrail, "Islands of Tractability for Parsimony Haplotyping," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 3, no. 3, pp. 303–311, 2006.
- [17] A. Tomescu, A. Kuosmanen, R. Rizzi, and V. Makinen, "A novel min-cost flow method for estimating transcript expression with RNA-Seq," *BMC Bioinformatics*, vol. 14, no. Suppl 5, p. S15, 2013. [Online]. Available: <http://www.biomedcentral.com/1471-2105/14/S5/S15>
- [18] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.
- [19] M. Garey and D. Johnson, "Complexity Results for Multiprocessor Scheduling under Resource Constraints," *SIAM Journal on Computing*, vol. 4, no. 4, pp. 397–411, 1975. [Online]. Available: <http://epubs.siam.org/doi/abs/10.1137/0204035>
- [20] B. Vatinlen, F. Chauvet, P. Chrtienne, and P. Mahey, "Simple bounds and greedy algorithms for decomposing a flow into a minimal set of paths," *European Journal of Operational Research*, vol. 185, no. 3, pp. 1390 – 1401, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0377221706006552>
- [21] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. New York, NY, USA: Springer-Verlag New York, Inc., 1985.